D354–D357 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj102

From genomics to chemical genomics: new developments in KEGG

Minoru Kanehisa[1,2,*], Susumu Goto[1], Masahiro Hattori[1], Kiyoko F. Aoki-Kinoshita[1], Masumi Itoh[1], Shuichi Kawashima[2], Toshiaki Katayama[2], Michihiro Araki[2] and Mika Hirakawa[1,3]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan, [2]Human Genome Center, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108-8639, Japan and [3]Institute for Bioinformatics Research and Development, Japan Science and Technology Agency, Chiyoda-ku, Tokyo 102-8666, Japan

Received September 14, 2005; Revised and Accepted October 17, 2005

ABSTRACT

The increasing amount of genomic and molecular information is the basis for understanding higher-order biological systems, such as the cell and the organism, and their interactions with the environment, as well as for medical, industrial and other practical applications. The KEGG resource (http://www.genome.jp/kegg/) provides a reference knowledge base for linking genomes to biological systems, categorized as building blocks in the genomic space (KEGG GENES) and the chemical space (KEGG LIGAND), and wiring diagrams of interaction networks and reaction networks (KEGG PATHWAY). A fourth component, KEGG BRITE, has been formally added to the KEGG suite of databases. This reflects our attempt to computerize functional interpretations as part of the pathway reconstruction process based on the hierarchically structured knowledge about the genomic, chemical and network spaces. In accordance with the new chemical genomics initiatives, the scope of KEGG LIGAND has been significantly expanded to cover both endogenous and exogenous molecules. Specifically, RPAIR contains curated chemical structure transformation patterns extracted from known enzymatic reactions, which would enable analysis of genome-environment interactions, such as the prediction of new reactions and new enzyme genes that would degrade new environmental compounds. Additionally, drug information is now stored separately and linked to new KEGG DRUG structure maps.

INTRODUCTION

While traditional genomics and other types of omics approaches have contributed to our knowledge on the genomic space of possible genes and proteins that make up the biological system, the new chemical genomics initiatives will give us a glimpse of the chemical space of possible chemical substances that exist as an interface between the biological world and the natural world. The KEGG database project was initiated in 1995, the last year of the first 5-year phase of the Japanese Human Genome Programme (1). After 10 years of development in parallel with the growing number of completely sequenced genomes and increased activities in post-genomic research, the KEGG project has entered a new phase in accordance with the chemical genomics initiatives.

KEGG is a database resource for understanding higher-order functions and utilities of the biological system, such as the cell or the organism, from genomic and molecular information. In fact, we consider KEGG as a computer representation of the biological system, consisting of building blocks and wiring diagrams, which can be used for modeling and simulation as well as for browsing and retrieval (2). Originally, the wiring diagrams involved endogenous molecules, both those that are directly encoded in the genome (proteins and RNAs) and those that are indirectly encoded through biosynthetic/biodegradation pathways (metabolites, glycans and so on). Now we are extending these wiring diagrams to include exogenous molecules. This will help understand interactions between the biological system and the natural environment, and would eventually lead to representation and reconstruction of another higher-level biological system, the biological world. Here we report new developments in KEGG towards this direction.

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp

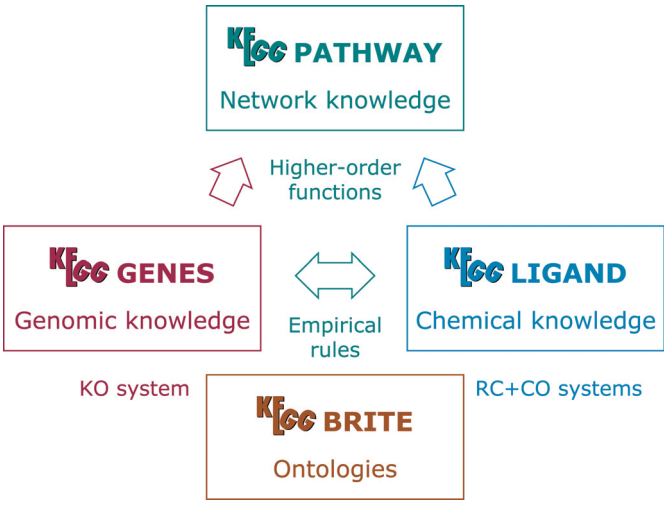

Figure 1. The overall architecture of KEGG now consisting of four main components. KEGG BRITE has been formally added to establish a logical foundation for inference of higher-order functions.

THE KEGG RESOURCE

Overview

KEGG consists of four main databases. As illustrated in Figure 1 they are categorized as building blocks in the genomic space (GENES databases) and the chemical space (LIGAND database), wiring diagrams in the network space (PATHWAY database) and ontologies for pathway reconstruction (BRITE database). BRITE had been a separate database for many years, but it was formally included in KEGG in release 34.0 (April 2005) to establish a logical foundation for the KEGG Project. The URLs for accessing KEGG are summarized in Table 1.

Biological systems are represented in KEGG by two types of graphs, called nested graphs and line graphs in theoretical computer science. The nested graph is a graph whose nodes can themselves be graphs. It is used for representing KEGG network hierarchy and for pathway reconstruction and functional inference. The line graph is a graph derived by interchanging nodes and edges of another graph. It represents the inherent complementarity of the metabolic pathway, which can be viewed either as a network of genes (enzymes) or as a network of compounds, meaning that one can be generated from the other by the line graph transformation. Thus, the line graph is the basis for integrated analysis of genomic and chemical information.

BRITE database

KEGG BRITE is a collection of hierarchies and binary relations with two inter-related objectives corresponding to the two types of graphs: to automate functional interpretations associated with the KEGG pathway reconstruction and to assist discovery of empirical rules involving genome-environment interactions. Currently, we focus on hierarchical structuring of our knowledge on functional aspects of the genomic and chemical spaces (Table 2), including the KEGG orthology (KO) system for ortholog/paralog gene groups, the reaction classification (RC) system for biochemical reactions, and other classifications for compounds and drugs tentatively called chemical ontology as shown in Figure 1. We plan to extend the KO system to include the definition of functional modules in the KEGG pathways and to develop ontologies for computational inference of higher-order functions.

Table 1. URLs for the KEGG resource

Database/content	URL
KEGG home page	www.genome.jp/kegg/
KEGG table of contents	www.genome.jp/kegg/kegg2.html
KEGG PATHWAY	www.genome.jp/kegg/pathway.html
KEGG GENES	www.genome.jp/kegg/genes.html
KEGG LIGAND	www.genome.jp/kegg/ligand.html
KEGG BRITE	www.genome.jp/kegg/brite.html
KGML	www.genome.jp/kegg/xml/
KEGG API	www.genome.jp/kegg/soap/
KEGG DRUG	www.genome.jp/kegg/drug/
KEGG GLYCAN	www.genome.jp/kegg/glycan/
KEGG REACTION	www.genome.jp/kegg/reaction/
KEGG EXPRESSION	www.genome.jp/kegg/expression/
KEGG ANNOTATION	www.genome.jp/kegg/kaas/
KegArray/KegDraw	www.genome.jp/download/
DBGET	www.genome.jp/dbget/
BLAST/FASTA	blast.genome.jp/
GenomeNet FTP	www.genome.jp/anonftp/
GenomeNet home page	www.genome.jp/

The current GenomeNet address 'www.genome.jp' is recommended, but the previous address 'www.genome.ad.jp' will still be made available.

Table 2. Functional hierarchies in KEGG BRITE

Network hierarchy
KO
Protein families
Enzymes
Transcription factors
Ribosome
Translation factors
ABC transporters
G-protein-coupled receptors
Ion channels
Cytokines
Cytokine receptors
Cell adhesion molecules (CAMs)
CAM ligands
CD molecules
Bacterial motility proteins
Compounds
Compounds with biological roles
Lipids
Phytochemical compounds
Compound interactions
Ion channel agonists/antagonists
Cytochrome P450 substrates
Drugs
Therapeutic category of drugs
Drug classification
Diseases
Disease genes, genomes and pathways
Organisms
KEGG organisms

As on September 12, 2005.

PATHWAY database

The KEGG PATHWAY database is a collection of manually drawn pathway maps for metabolism, genetic information

processing, environmental information processing such as signal transduction, various other cellular processes and human diseases. During the past 2 years we have significantly increased the number of pathway maps for regulatory pathways including signal transduction, ligand–receptor interaction and cell communication, all based on extensive survey of published literature. For metabolic pathways we created two new sections, 'Glycan Biosynthesis and Metabolism' and 'Biosynthesis of Polyketides and Nonribosomal Peptides'. The XML version of the pathway maps is available for both metabolic and regulatory pathways. These KEGG Markup Language (KGML) files provide graph information that can be used to computationally reproduce and manipulate KEGG pathway maps.

GENES database

The KEGG GENES database is a collection of gene catalogs for all complete genomes and some partial genomes (31 eukaryotes, 235 bacteriaand 23 archaea as of September 12, 2005), generated from publicly available resources, mostly NCBI RefSeq (3). All genomes in KEGG GENES are subject to SSDB computation and given manual KO assignments as described below. There are auxiliary collections of gene catalogs: DGENES for draft genomes (21 eukaryotes) and EGENES for expressed sequence tag consensus contigs (25 plants). These are meant to supplement the repertoire of KEGG organisms, and all are given automatic KO assignments using GENES as a reference dataset. Each GENES entry contains cross-reference information to outside databases, including NCBI gi numbers, Entrez Gene IDs and UniProt accession numbers. Starting with KEGG release 37.0 (January 2006) automatic ID conversion is implemented enabling use of such outside identifiers to access KEGG GENES and then the other KEGG databases.

KEGG orthology

There is a total of over one million genes in KEGG GENES, representing a tiny, but well-characterized part of the genomic space that makes up the biological world. From this part we organize knowledge about orthologous genes and paralogous genes, which, we hope, can be generalized for understanding the entire genomic space. This knowledge is stored in the KO system, a pathway-based classification of orthologous genes, including orthologous relationships of paralogous gene groups. The KO identifier, or the K number, is a common identifier for linking genomic information in the GENES database with network information in the PATHWAY database. The pathway nodes represented by rectangles in the KEGG reference pathway maps are given KO identifiers, so that organism-specific pathways can be computationally generated once each genome is annotated with KO's. This annotation or the KO assignment is done manually for KEGG GENES with the help of the GFIT tool using best-hit relations in pairwise genome comparisons stored in the SSDB database (4).

Because the number of ortholog groups that can be linked to pathways is limited, we have introduced two additional ways to define KO's. One is to use COG (5) to cover a broad-range of possible ortholog groups. The other is to rely on experts' classifications of protein families, which tend to be more functionally oriented resulting in narrowly defined KO's. A growing number of protein families are being added to the KO system, and they are shown in separate hierarchies different from the KEGG network hierarchy. The KO system can be best viewed from the KEGG BRITE database (Table 2).

LIGAND database

Originally, the LIGAND database consisted of just two components: ENZYME for enzyme nomenclature and COMPOUND for chemical compound structures (6). It later successively included additional components: REACTION for chemical reaction formulas, GLYCAN for glycan structures, RPAIR for reactant pair transformation patterns and DRUG for drug information. This expansion of the LIGAND collection represents our expanded efforts for understanding the chemical space that is part of the biological world.

The KEGG DRUG database is a new addition from KEGG release 36.1 (December 2005). It contains chemical structures and additional information such as therapeutic categories and target molecules. A most unique feature of KEGG DRUG is a collection of drug structure maps, which graphically illustrate, in a manner similar to KEGG pathway maps, our knowledge on groups of chemical structural patterns, therapeutic categories, their relationships and the chronology of drug development if known.

Reaction classification

The RC system in the chemical space is a counterpart of the KO system in the genomic space (Figure 1). It represents our attempt to organize knowledge on chemical reactions by categorizing chemical structure transformation patterns. The REACTION database contains individual reaction formulas taken from the ENZYME database. Each reaction formula is split into a set of substrate-product pairs, and the chemical structure comparison program SIMCOMP is applied to obtain an optimal alignment. This comparison is based on atom typing, which is the conversion of regular atomic (C, N, O, S, P and so on) representation to what we call KCF representation that consists of 68 atom types distinguishing functional groups and atomic environments (7). The chemical structure alignment generated by SIMCOMP is used to define the R atom for the reaction center, the D atom(s) for adjacent atom(s) in the mismatched region and the M atom(s) for adjacent atom(s) in the matched region (8). This is first done computationally and is followed by extensive manual curation.

The RPAIR database is still under development, but it is the basis for the RC system categorizing curated RDM patterns. Since an enzymatic reaction usually involves multiple substrates and products, one EC number corresponds to a combination of RDM patterns. The RC system has enabled automatic assignment of EC numbers from a set of substrate and product structures (8) and will further enable exploration of unknown reactions by generating plausible combinations of RDM patterns, which may then be related to possible paralogs of enzyme genes.

Glycosyltransferase reactions

Functional glycomics has been a most successful area for integrated analysis of genomic and chemical information (9). The carbohydrate sequence of glycans is determined by a specific set of biosynthetic reactions catalyzed by different

types of glycosyltransferases. Thus, once we know the repertoire of glycosyltransferases in the genome or in the transcriptome, it should in principle be possible to predict the repertoire of glycan structures. Conversely, the knowledge about glycan structures can be used to search and annotate new glycosyltransferases. Composite Structure Map in KEGG GLYCAN is a tool for converting genomic or transcriptomic data to glycan structure variations based on a curated set of known glycosyltransferase reactions.

ACCESSING KEGG

Web and FTP

KEGG is the major component of the Japanese GenomeNet, which is served by the Kyoto University Bioinformatics Center. The other GenomeNet services including DBGET and BLAST/FASTA searches are now primarily developed and used to support KEGG. The official URL for GenomeNet has been modified to http://www.genome.jp/, but the former URL http://www.genome.ad.jp/ will still be made available (Table 1). To download the KEGG data, academic users may use the GenomeNet FTP site.

KEGG API

The KEGG API service has become an increasingly popular mode of access. It is the SOAP/WSDL interface to KEGG, enabling users to write their own programs to access, customize and utilize KEGG.

KegArray and KegDraw

KegArray and KegDraw are standalone Java applications that make use of the KEGG resources. KegArray is for microarray data analysis in conjunction with KEGG pathways and genomes. KegDraw is for drawing glycan structures and chemical compound structures, which can then be used to query against KEGG and PubChem databases. Both are freely available to academic and non-academic users.

ACKNOWLEDGEMENTS

The KEGG project is supported by the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, the 21st Century COE program 'Genome Science', and a grant-in-aid for scientific research on the priority area from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. Funding to pay the Open Access publication charges for this article was provided by the grant-in-aid for scientific research.

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M. (1997) A database for post-genome analysis. *Trends Genet.*, **13**, 375–376.
2. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
4. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
5. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
6. Goto,S., Nishioka,T. and Kanehisa,M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.
7. Hattori,M., Okuno,Y., Goto,S. and Kanehisa,M. (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
8. Kotera,M., Okuno,Y., Hattori,M., Goto,S. and Kanehisa,M. (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J. Am. Chem. Soc.*, **126**, 16487–16498.
9. Hashimoto,K., Goto,S., Kawano,S., Aoki-Kinoshita,K.F., Ueda,N., Hamajima,M., Kawasaki,T. and Kanehisa,M. (2005) KEGG as a glycome informatics resource. *Glycobiology*, in press.